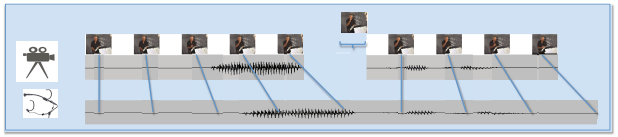
Audio Synchronisation with a Tunnel Matrix for Time Series and Dynamic Programming

Jan Gorisch [1 2] and Laurent Prévot [1]

[1] Laboratoire Parole et Langage & CNRS, AMU, France; [2] Division of Linguistics and Multilingual Studies, NTU, Singapore

Introduction

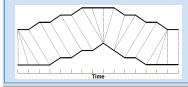
- Multimodal analyses of spoken interaction require high quality recordings of audio, video, and potentially other signals.
- Sometimes these signals are out of sync and need to be resynchronised in order to render them analysable.
- Sources of error are 1) clocking error between recording devices, or 2) image loss.



Current approaches

Hand correction: inspection of audio and video by ear and eye and cut video accordingly. Very time-consuming.

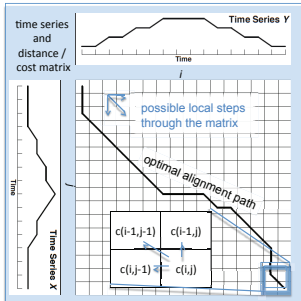
Additional equipment: a disk on a turn table with scratch and visible marker is recorded on a separate audio channel and camera. Information is used for synchronisation [1]. Equipment might also fail and complicate the recording setup.



Dynamic programming: align signals from two audio recordings. This has been used successfully for short audio and video clips of approximately 3 minutes length [2]. This method is basically adopted here.

Background on Dynamic Programming (DP)

The aim of DP is to describe how two time series that share some characteristics, but that do not match in the time domain, need to be manipulated in order to match each other. An alignment path shows necessary insertions and deletions.



example borrowed from Salvador & Chan (2004)

Two time series *X* and *Y* need to be aligned.

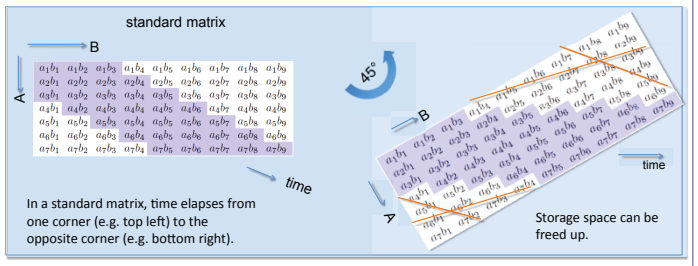
1. Distance matrix *D*: calculate the pair-wise distances: $d_{i,j} = |X_{i,j} - Y_{i,j}|$
2. Cost matrix *C*: accumulate the distances and the minimum cost-up-to-this-point looking at each possible predecessor:
$$c_{i,j} = \min \begin{bmatrix} c_{i,j-1} + d_{i,j} \\ c_{i-1,j} + d_{i,j} \\ c_{i-1,j-1} + 2d_{i,j} \end{bmatrix}$$
3. Backtrack matrix: remember at each step where the path came from: (1) horizontal, (2) vertical, (3) diagonal
4. Alignment path: once the final point is reached in $C_{I,J}$, the optimal alignment path is achieved through backtracking

Problems of standard DP

Storage space: the alignment of audio recordings of the length of 30min (1,800s), even using a relatively sparse sampling rate of 25Hz (as standard video frames), results in a distance matrix of (25 * 1,800)², that are 2,025,000,000 value pairs.
Computational cost: the evaluation of the optimal alignment path is extremely costly. It can be reduced with e.g. the Sakoe Chiba Band [3], but the entire matrix still needs to be stored. A sparse matrix [4], could be used for such a band, but DP has not been demonstrated on sparse matrices, yet.
Solution: in order to save space and computation power to solve the desynchronisation problem at hand, we suggest a new structure for the values of the distance matrix. It is organised in a "tunnel" form.

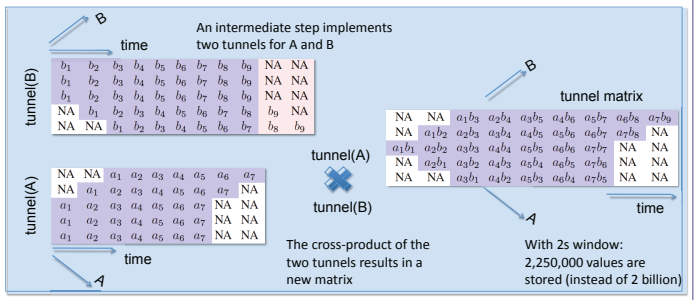
Implementation of the Tunnel Matrix

The aim is to rotate the Sakoe Chiba Band (colored area) from a standard distance matrix (left) around the diagonal by – 45° in order to free up space. Time elapses then from left to right.



In a standard matrix, time elapses from one corner (e.g. top left) to the opposite corner (e.g. bottom right).

Storage space can be freed up.



An intermediate step implements two tunnels for A and B

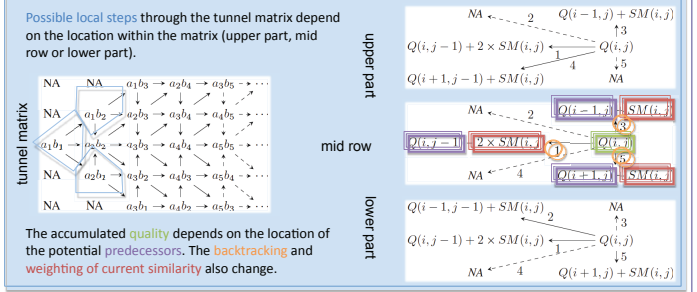
The cross-product of the two tunnels results in a new matrix

With 2s window: 2,250,000 values are stored (instead of 2 billion)

Tunnel DP-algorithm

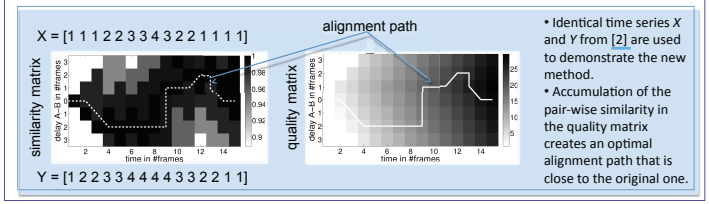
The tunnel matrix requires an adaptation of the DP algorithm to cope with the new structure.

Possible local steps through the tunnel matrix depend on the location within the matrix (upper part, mid row or lower part).



The accumulated quality depends on the location of the potential predecessors. The backtracking and weighting of current similarity also change.

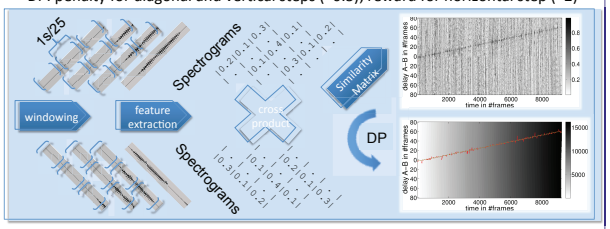
Results on artificial data



- Identical time series *X* and *Y* from [2] are used to demonstrate the new method.
- Accumulation of the pair-wise similarity in the quality matrix creates an optimal alignment path that is close to the original one.

Results on real data

- mono from camera stereo mic. and mono mix-down from both headset mics.
- windowing (Hamming) according to expected accuracy of output (25fps)
- extraction of acoustic representation as spectrograms with 265 frequency points
- tunnel A and B made of spectrograms; then cross-multiply → similarity matrix
- DP: penalty for diagonal and vertical steps (*0.5); reward for horizontal step (*2)



Output after DP as alignment vector V_{align}

- Apply median smoothing (window length of 399 frames) to alignment path.
- Take the derivative of smoothed path.

V_{align}=[11111110000010000010000…0001000…0000000-10000000]

- Use this information to add or delete images in the video.

1 -1

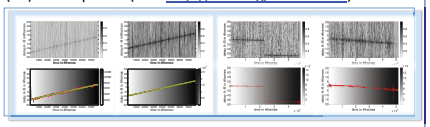
Resynchronisation

- Extract images from video: script calling with FFmpeg (ffmpeg.org).
- Copy and delete image files: script using the output vector [11001000…].
- Re-order image files: rename: 00001.jpg – 00002.jpg – 00003.jpg …
- Re-combine film from images: script calling ffmpeg.

Evaluation

- Most of the audio-visual Aix MapTask corpus (MTX) (2h) [5] was repaired (see output vectors, similarity and quality matrices: http://sldr.org/sldr000875).
- The entire DVD corpus (4h) was repaired (see: http://sldr.org/sldr000891).

Examples:
MTX (left) DVD (right)
spk. AD + CR spk. AM + GM



Future work

- Expand to other signals (EEG, EMA, acoustic-prosodic features, etc.).
- Apply to interactional alignment, done by conversational participants on a turn-by-turn basis [6], or on a larger time-scale, as in convergence [7].

References

[1] Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House, "Spontal: A Swedish spontaneous dialogue corpus," in *Proceedings of LREC'10*, Malta, 2010.
[2] Stan Salvador and Philip Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," in *The Third SIGKDD Workshop on Mining Temporal and Sequential Data*, Seattle, WA, USA, 2004, pp. 70–80.
[3] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
[4] John R. Gilbert, Cleve Moler, and Robert Schreiber, "Sparse matrices in matlab: Design and implementation," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, pp. 333–356, 1992.
[5] Jan Gorisch, Corine Astésano, Ellen Gurman Bard, Brigitte Bigi, and Laurent Prévot, "Aix Map Task corpus: the French multimodal corpus of task-oriented dialogue," in *Proceedings of LREC'14*, Reykjavik, Iceland, 2014.
[6] Jan Gorisch, Bill Wells, and Guy J. Brown, "Pitch contour matching and interactional alignment across turns: an acoustic investigation," *Language and Speech*, vol. 55, no. 1, pp. 57–76, 2012.
[7] Spyros Kousidis, David Dorran, Yi Wang, Brian Vaughan, Charlie Cullen, Dermot Campbell, Ciaran McDonnell, and Eugene Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues," in *Proceedings of Interspeech 2008*, Brisbane, Australia, 2008.

Acknowledgments

This work is supported by the Agence Nationale de la Recherche, grant number ANR-12-JCJC-JSH2-006-01. The first author also benefits from a mobility allowance from the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7.

The author has requested enhancement of the downloaded file. All in-text references underlined in blue are linked to publications on ResearchGate.